This is the author's version. To cite this article:

Vieira, L.N. (2017) Cognitive effort and different task foci in post-editing of machine translation: a think-aloud study. *Across Languages and Cultures* 18(1): 79-105. doi: 10.1556/084.2017.18.1.4

COGNITIVE EFFORT AND DIFFERENT TASK FOCI IN POST-EDITING OF MACHINE TRANSLATION: A THINK-ALOUD STUDY

LUCAS NUNES VIEIRA

University of Bristol
School of Modern Languages
17 Woodland Road
Bristol BS8 1TE United Kingdom
Phone: +44 (0)117 9288610
E-mail: l.nunesvieira@bristol.ac.uk

Abstract: Post-editing of machine translation is gaining popularity as a solution to the ever-increasing demands placed on human translators. There has been a great deal of research in this area aimed at determining the feasibility of post-editing, and at predicting post-editing effort based on source text features and machine translation errors. However, considerably less is known about the mental workings of post-editing and post-editors' decision-making or, in particular, the relationship between post-editing effort and different mental processes. This paper investigates these issues by analysing data from a think-aloud study through the lens of eye movements and subjective ratings obtained in a separate task. The results show that mental processes associated with grammar and lexis are significantly associated with cognitive effort in post-editing. This association was not observed for other aspects of the task concerning, for example, discourse or the real-world use of the text. In addition, it was noted that lexical issues are linked to long sequences of thought processes. The paper shows that lexis plays a central role in post-editing, and argues that more emphasis should be placed on this issue in future research and in post-editor training.

Keywords: post-editing, machine translation, cognitive effort, eye tracking, think-aloud protocols

1. INTRODUCTION

Post-editing machine translation (MT) output has become extremely popular as a less costly and potentially more effective alternative to traditional translation (see Green, Heer, and Manning 2013). Post-editing has also attracted attention in academia, both as a translation modality in its own right (e.g. Krings 2001) and as a strategy for automatically estimating

MT quality (Specia 2011). Much research has focused on predicting effort and examining the feasibility of post-editing in terms of translating productivity and translation quality. However, the nature of mental processes in post-editing and the relationship between these processes and *cognitive* effort (a sub-type of overall post-editing effort – see section 2.1) have received considerably less attention. Information of this kind is able not only to enhance the general understanding of post-editing, but also to help characterise the notoriously elusive concept of cognitive effort in this specific context. The think-aloud method, whereby participants verbalise their thoughts during a task (see Ericsson and Simon 1980), was used here to investigate these issues.

Data from a task carried out under the think-aloud condition, henceforth 'the think-aloud task', was used to cast light on the different aspects of the task participants focused on, such as lexis, grammar, and readership-specific issues. These aspects are referred to here as 'task foci'. They consist of issues that participants think of and/or address, as indicated by the think-aloud data or edits in the MT output. Think-aloud protocols (TAPs) were analysed in their own right as well as in the light of data obtained in a separate task, where eye tracking and subjective ratings were used to estimate cognitive effort – henceforth 'the eye-tracking task'. The think-aloud and eye-tracking tasks were carried out by different, but comparable, participant samples in the context of a larger project (see Vieira 2016). The same texts were used in the two tasks, which made it possible to identify cognitively demanding text passages based on the eye-tracking task and then check to see, based on the think-aloud data, what participants think and do when post-editing these passages.

This mixed-method setup involving TAPs and eye movements, which to the author's knowledge has not been attempted in a post-editing study before, allowed cognitive effort to be investigated while compensating for the interference posed by eye tracking and the think-aloud condition. While eye tracking normally entails constraints in the task design and editing interface, the think-aloud condition is deemed to interfere with linguistic tasks (Krings 2001; Jakobsen 2003), which casts doubt on whether it is a suitable method for investigating effort. This design follows the principle of triangulation (see e.g. Creswell 2009:14-16), whereby inevitable weaknesses inherent to different methods are compensated through a combination of data sources. In the present study, data from the eye-tracking task allows for an independent estimation of cognitive effort, serving as a framework for the TAPs analysis. The think-aloud data, in turn, allows qualitative details of post-editors' mental processes to be investigated, which would not be possible with the use of eye tracking alone.

In what follows, the concept of cognitive effort is briefly outlined in section 2, together with a review of previous research. The study's methodology is described in section 3. Data-processing steps are explained in section 4, and results are reported in section 5. Concluding remarks are presented in section 6.

2. BACKGROUND AND PREVIOUS RESEARCH

2.1. Cognitive Effort

In cogntive psychology, cognitive effort is defined as 'the amount of the available processing capacity of the limited-capacity central processor utilized in performing an information-processing task' (Tyler et al. 1979:608). This definition draws on previous research by Moray (1967), who proposed that the brain works as a processor that establishes the quantity of mental resources to be allocated to a task based upon characteristics of the task itself. The allocated resources are what Tyler et al. call *effort*.

In the specific context of post-editing, Krings (2001:179) defines three types of effort: the *technical* effort posed by merely mechanical operations, the *cognitive* (i.e. mental) effort required by the task, which consists of the 'type and extent of […] cognitive processes' (ibid. 179) that take place, and *temporal* effort, i.e. post-editing time. Cognitive effort is a construct that cannot be quantified based on direct measures – as can temporal effort, for example – so any investigation of cognitive effort will require the use of indirect parameters.

In line with previous research in post-editing, the present study uses eye movements and subjective ratings to estimate cognitive effort (see e.g. O'Brien 2011; Koponen 2012). The use of eye tracking as a cognitive method is based on the eye-mind and immediacy assumptions, according to which the mind necessarily processes the information received by the eyes during reading, and does so without delay (see Just and Carpenter 1980). This method has been used extensively in previous studies on translation and post-editing, including those stemming from the CASMACAT project (Koehn et al. 2015). As in recent post-editing research (e.g. Alves et al. 2016), fixation count and the average duration of eye fixations were the specific eye-tracking metrics used here. The use of subjective ratings to estimate effort, in turn, is based on the assumption that individuals are able to report on cognitive (or mental) effort expenditure in terms of numerical scores (see O'Donnel and Eggemeier 1986; Paas 1992). Paas (1992) adapted a scale used in previous research for this purpose and proposed a nine-point self-report scale that can be used for 'translating the

perceived amount of mental effort into a numerical value' (Paas 1992:430). Paas's scale was used in the present study. This scale is a well-established instrument to measure the mental effort invested in a task (see Paas et al. 2003:68).[1]

2.2. Related Work

To the present author's knowledge, Krings (2001) carried out the only investigation in MT post-editing to date based on TAPs collected concurrently to the post-editing task. Krings aimed to investigate the effort required by traditional translation and by post-editing, as well as the feasibility of carrying out post-editing without looking at the source text (ST). In addressing these aims, Krings provided a description of post-editing's mental processes based on fine-grained operational actions; for example, 'machine/read' (reading the MT output) or 'target/mon/compare/ST-MT' (comparing the raw MT output with the ST) (2001:514ff). The present study, by contrast, is interested in the thought processes that underlie post-editors' attentional foci, i.e. the different aspects of the task post-editors focus on when they make decisions. In describing the process of mentally parsing the text, Krings (2001) distinguishes between different levels of linguistic analysis, such as 'syntax' and 'pragmatics'. However, Krings's focus is on the incidence of these levels within post-editing processes related to the ST, the MT, and the TT (i.e. the emerging edited text), rather than on how these levels relate to cognitive effort, a question that is addressed in the present study. In addition, as previously mentioned, the think-aloud method might interfere with linguistic tasks and represent in itself a source of effort (Jakobsen 2003). This is a downside of Krings's design which is combatted by the mixed-method approach adopted here.

A number of studies exploit linguistic aspects of the ST and/or MT output as potential indices of effort in post-editing (e.g. Aziz, Koponen and Specia 2014; Vieira 2014; O'Brien 2011). It should be noted that these studies explore the connection of effort with quantitative textual features (e.g. the incidence of different part-of-speech categories) and not with mental processes, so their objective is different from the one addressed here.

In an earlier study, Temnikova (2010) developed a typology of MT errors ranked according to the levels of cognitive effort the errors were expected to require in post-editing. She suggested that problems stretching across longer textual spans, such as those of a syntactic nature, should be regarded as more cognitively demanding than local errors, such as missing words. Koponen et al. (2012) built on this typology and measured the post-editing time required by sentences containing different error categories among those proposed by

Temnikova. Koponen et al. found that long-span errors are associated with more post-editing time per target word, a temporal measure that they put forth as an indicator of cognitive effort. Lacruz, Denkowski, and Lavie (2014) used pause-to-word ratios to estimate cognitive effort, and found that MT errors involving mistranslation and omission/addition are especially cognitively demanding. With regard to these studies, it should be noted that, as a professional translating modality in its own right, post-editing would be expected to involve mental processes associated with a number of other factors in addition to the MT output, such as the intended readership and the real-world use of the text. Aspects of this kind are jointly analysed in the present study in what, to the knowledge of the author, is the first time that these factors are contrasted with the levels of cognitive effort expended by post-editors.

3. METHODOLOGY[2]

3.1 Post-Editing Tasks

Participants post-edited extracts of two news articles taken from the *newstest2013* corpus. This corpus includes STs, raw MT outputs and reference human translations resulting from the 2013 edition of the Workshop on Statistical Machine Translation.[3] French-to-English was the language combination adopted for the study. The source articles were about prostate cancer[4] and the 2012 United States elections,[5] respectively. As previously mentioned, participants in the eye-tracking and think-aloud tasks post-edited the same texts.

To investigate the post-editing process in a range of conditions, the study involved MT outputs of a range of quality levels. The automatic translation evaluation system Meteor (Denkowski and Lavie 2011), which assesses the similarity between machine-translated sentences and corresponding human reference translations, was used to evaluate the quality of potential MT sentences to be used in the study and ensure that sentences at different quality levels were selected. Meteor scores range from 0 (no similarity between MT and human reference translation) to 1 (perfect match). The MT sentences were taken from the *newstest2013* corpus as well as from online MT systems,[6] which helped to increase variability in MT quality. In total, participants worked through 1037 source words. A sample of 41 sentences (844 source words) was selected for the analysis of cognitive effort data. Titles and sentences for which reference translations were not available were not considered in analyses involving cognitive effort to avoid acclimatisation effects or because Meteor scores could not be computed (see Vieira 2014:193-194). The sample had Meteor scores ranging between 0.14 and 1.[7]

In the eye-tracking task, participants post-edited the texts in PET (Aziz, Castilho, and Specia 2012) in document order. One sentence was shown on screen at a time, which was necessary to make sure that the eye-tracking data was of good quality. Participants provided subjective ratings on cognitive effort immediately after editing each sentence, based on the scale described in section 2.1. This scale was set up in PET's interface; it varies between 1, 'very, very low mental effort' and 9, 'very, very high mental effort' (Paas 1992). Subjective ratings were only collected in the eye-tracking task.

In the think-aloud task, participants post-edited the texts in Translog-II (Carl 2012), having access to the entire texts on the screen. Translog-II produces linear key-logging reports that could be used to support the TAPs analysis, so this tool was deemed a more suitable editing interface for the think-aloud task.

Both tasks included a warm-up phase that served to acquaint participants with the set-up (e.g. the editing tools and the think-aloud condition, in the think-aloud task). The order of presentation of the texts was alternated between participants in both tasks. Participants were told to aim for post-edited texts that would be suitable for publication in an English-speaking context. They were asked to carry out the tasks as fast as possible, but no time limit was imposed. The configuration of Translog-II's interface is presented in Figure 1. The ST was displayed on the left and the machine-translated text on the right.

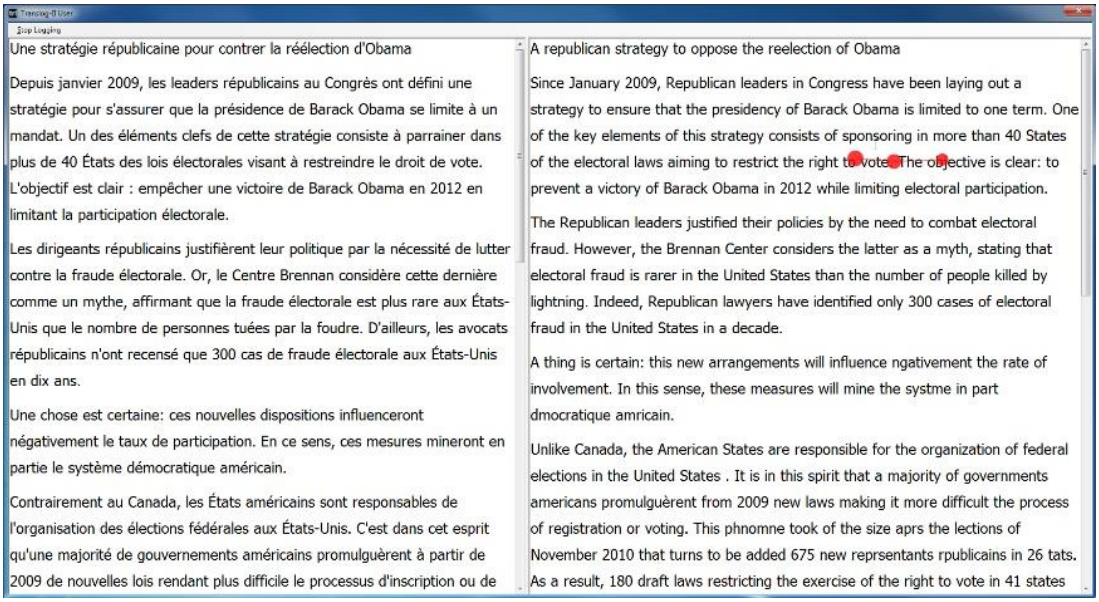

Figure 1
Translog-II interface, as used in the think-aloud task – blobs indicate eye fixations

Participants' eye movements and verbalisations (in the think-aloud task) were recorded with Tobii Studio, which was also used to record the screen. Tobii X120 was the

eye tracker used.[8] The tasks were conducted on site, at Newcastle University. Eye-tracking data was collected in both tasks, but it was not used for quantitative data analysis in the think-aloud task, as the think-aloud condition could have altered participants' eye-movement behaviour. In the eye-tracking task, gaze data was obtained with Tobii Studio by demarcating as an 'area of interest' the area on screen where the sentence pairs (i.e. ST and MT/TT) were displayed, which allowed the data corresponding to this area to be collected and processed. Because of constraints imposed by the collection of gaze data, participants were not permitted to consult external sources in the eye-tracking task (see e.g. Hvelplund 2011:86-87). This restriction was maintained in the think-aloud task for consistency. Prior to post-editing each news article, in both tasks participants were nevertheless asked to read a text that briefly explained the articles' subject matter. This was deemed desirable as a way of reconciling potential discrepancies in participants' subject-matter knowledge and compensating for the restriction on the use of external sources. Participants were asked to retain MT suggestions if they did not know or could not infer the meaning of ST words.

The quality of the post-edited translations was assessed in terms of fluency and adequacy (see Vieira 2016), but these results are not reported here so as not to deviate from the objective of this paper: investigating associations between cognitive effort and different task foci in post-editing.

3.2 Participants

All participants in the study were native speakers of English. The sample included professional translators, translation students, and non-professionals who were starting to work as translators or who had an educational background in translation.[9] Participants were sampled from the student population at Newcastle University and from networks of professional translators based in the North East of England. Nineteen participants carried out the eye-tracking task. Other ten participants carried out the think-aloud task, but one of these was excluded from the sample because of a difficulty to think aloud. Participants' average age was 33.4 (SD = 15.7).

After post-editing the texts, participants' level of proficiency in French was measured with a vocabulary task that is often used as a placement test (see Meara and Buxton 1987; Read 2007). To control for participants' attitude to MT (see de Almeida 2013), they also rated their opinion on the use of MT in human translation workflows, choosing a level between 1 (negative) and 5 (positive).

Table 1 presents participants' profile. Care was taken in attempting to ensure that the two task samples had a comparable profile. Small differences can nevertheless be observed between participants in the eye-tracking and think-aloud samples: those in the think-aloud sample had a higher level of professional experience and a higher level of proficiency in French. However, Wilcox-Mann Whitney tests showed that these differences were not significant.[10]

Table 1
Participants' profile – per-participant values for the think-aloud task and means for both tasks.

Participants	**FR Vocab (0-100)**	**Experience (in years)**	**Attitude (1-5)**
P20	95	13	2
P21	75	0	3
P22	90	0	3
P23	10	0	5
P24	97	0.6	3
P25	98	28	2
P26	93	10	4
P27	98	0	5
P28	78	0.1	3
Think-aloud task mean (N = 9)	*81.5*	*5.7*	*3.1*
Eye-tracking task mean (N = 19)	*70.4*	*2.1*	*3.4*

It should be noted that some participants had a very basic knowledge of French (e.g. P23). This was not considered problematic, as French proficiency was controlled for in the analysis, and further tests were carried out to see if this variable affected the results (see section 5).

4. DATA PROCESSING AND CODING

After transcribing the data, the TAPs were segmented and coded (see Sun 2011:943; Krings 2001:309-310). These procedures are described below in sections 4.1 and 4.2.

4.1 Task Phases and TAPs Segmentation

Figure 2 shows total post-editing duration per text in the think-aloud task for all participants. It was noted that the tasks were divided into three phases, similarly to a pattern observed by Carl, Kay, and Jensen (2010). In the context of traditional translation, Carl, Kay, and Jensen call these phases 'gisting', 'drafting', and 'post-editing', i.e. skimming the ST, typing a first

draft of the translation, and revising the draft, respectively. In the present study, only P26 went through an initial gisting phase. In both post-editing sessions this participant skimmed through the MT output before performing any edits, making statements such as 'I'm going to start by reading the English text'.

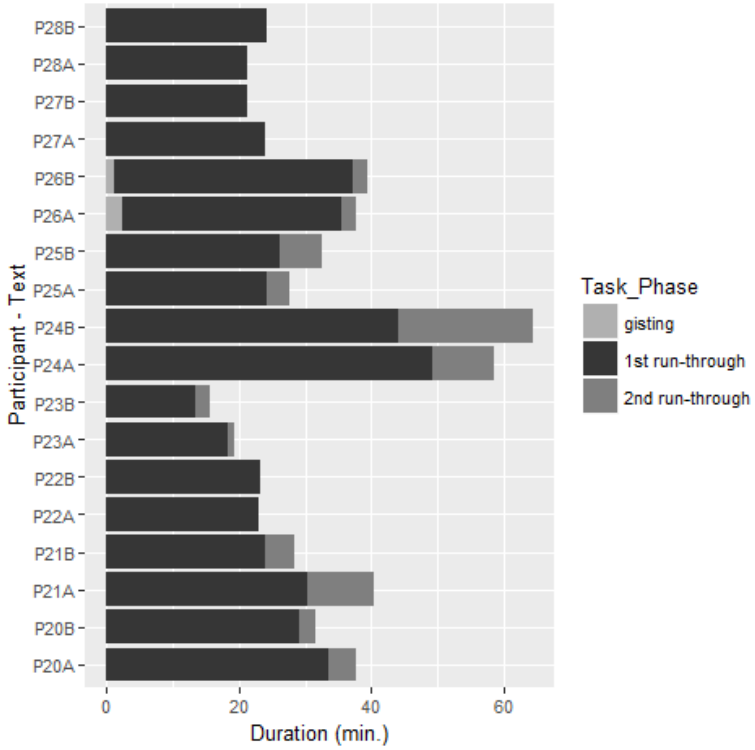

Figure 2
Think-aloud task durations (x-axis) per text, per participant (y-axis)

Most of the other participants carried out the tasks in two stages: a first run-through where they performed the majority of edits, and a second run-through where they revised their work. Even though they often backtracked while still in the first run-through, the beginning of a second run-through was clear when participants returned to the beginning of the text and started again, often making statements such as 'OK, I'm going to have a final read-through'. Seven participants (out of 9) presented this behaviour. Since there were participants who did not have a gisting phase or a second run-through, only TAPs corresponding to the first run-through were considered in the quantitative analysis presented in Section 5.

The transcribed data was segmented into coding units according to rules proposed by Krings (2001:309-310). These rules were also designed for post-editing, so they seemed particularly suitable for the present study. Pauses were the main segmentation criterion followed by Krings. Based on previous research, Krings regarded pauses of at least one

second as a coding unit divider (2001:304). He proposes a number of other dividers in addition to pauses, including shifts of attention to/from the 'object language' (i.e. the ST, MT output or TT, as opposed to participants' comments), shifts between different editing solutions, and shifts to/from physical writing events. Writing events were only regarded as a separate coding unit in the present study when they were not verbalised nor accompanied by any comment. Krings also proposed a final fusion rule whereby pauses are disregarded if verbalisation units are homogenously linked in the same proposition. This rule too was followed here.

(1)
U315 I wonder I think *republican* will have a capital *R*
U316 because it's a political party
U317 uh not ~~*defined*~~ *a strategy*
U318 maybe ***drawn up***
U319 or *set out*
U320 anyway
U321 *drawn up a strategy to ensure that the presidency*
U322 *a mandate*
U323 no
U324 *one*
U325 *a single mandate*
U326 I don't know if you'd really say that
U327 *a* ***single term*** [~~*mandate*~~]

KEY:
bold: insertions; ~~strikethrough~~: deletions; *italics*: ST or MT or TT; [ ]: physical writing processes not spoken out loud or comments added by the researcher; underline: participant's emphasis

An extract of the TAPs produced by P20 is provided in example (1), where each line is a coding unit (U).

4.2 Coding

The study's coding procedure consisted of a combination of inductive and deductive strategies: while the coding categories were motivated by the research question, they were also influenced by the data itself, and accommodated any unforeseen phenomena.

First, the present author coded all the data. An external coder was subsequently involved in the study to fine-tune the coding scheme and to measure inter-coder reliability based on a random sample with TAP sequences (see Section 5.3) amounting to a total of 100 coding units. The external coder was a native English speaker and had a PhD in French. After preliminarily checking inter-coder reliability and discussing any discrepancies, a Cohen's Kappa of 0.63 was ultimately obtained, based on a second random sample. According to

Landis and Koch (1977), this represents ‘substantial’ agreement, so this result was deemed adequate for the study.

The coding categories were divided into two groups: specific task foci and non-specific task foci. Specific task foci were those that corresponded to a specific linguistic aspect of the task. Non-specific task foci did not correspond to any explicit linguistic issues. The coding categories and examples are presented below.

4.2.1 Specific Task Foci

- Lexis

This category had TAP units concerned with lexical meaning, including issues related to collocations and fixed expressions. These units usually involved content words. E.g.:

(2) U560 it's kind of a euphemism that has not been translated suitably
U1478 I'm trying to think what the word is in English for that

- Grammar/Syntax

Units in this category (henceforth ‘Grammar’) were related to aspects of grammar or syntax, such as the use of the passive/active voice, number agreement, verb tense, etc. These units often involved function words. E.g.:

(3) U395 ***These*** [~~*This*~~] *new arrangements will influence*
U24 ah, word order is wrong [~~*States*~~]

- Discourse

These units related to aspects around punctuation, document consistency, the links between sentences, and other issues concerned with the texts’ overall coherence. E.g.:

(4) U535 I wonder how those two sentences were put together
U22 too many commas?

- Style

These units related to aspects concerning the texts’ flow/style. This category was only used when the issue in question did not fit the Grammar category. This was the case with issues concerning sentence length, wordiness, repetition, the addition of words to improve the flow

of the text, and word order shifts that did not involve grammatical modifications (e.g. ‘therefore recommend’ rather than ‘recommend therefore’). E.g.:

(5) U491 I think ~~*such a*~~ *requirement* is not necessary
U2066 that sounds too wordy

- Translation Context/World Knowledge

Units in this category (henceforth ‘Knowledge’) were related to the translation’s real-world context. These units involved aspects such as readership, genre, source/target cultures, the real-world use of the text, subject matter, intertextuality, etc. E.g.:

(6) U497 I'd probably check whether *constitutionality* is used in America
U15 *prostate cancer* is better for the headline

- Orthography/Capitalisation/Typography

Units in this category (henceforth ‘Orthography’) related to aspects such as spelling, capitalisation, and the number of spaces after punctuation. E.g.:

(7) U446 I think it's only one space after the full stop
U146 I presume *African* and *American* is they're [sic] both capitalised

4.3.2 Non-Specific Task Foci

- Non-Specific Reading/Evaluation

Three modes of reading were observed in the data: (*i*) an initial reading mode where text segments were put into working memory for mental processing, (*ii*) reading events that were related to specific editing issues, and (*iii*) reading events aimed at revising any modifications or checking if there were still any problems that needed to be addressed. Reading mode *ii* was coded with a corresponding specific task foci category, as a motive for the reading event was explicit in these cases. Conversely, reading modes *i* and *iii* were coded with the present category (henceforth ‘Reading’), since the motive for these reading events was either neutral or unspecified. This also applies to positive/negative evaluations of the text that did not have an explicit motivation. E.g.:

(8) U81 and I think the last sentence is OK
U4957 *because the cancer is not aggressive and does not threaten their lives* [initial reading]

- Procedural

Units in this category involved procedural aspects of the task, such as when participants mentioned what they were about to do. E.g.:

(9) U1348 I'm going to read this in French
U2035 OK I'll come back to that

An 'Undefined' coding category was used for any special cases that did not fit the categories described above. For consistency, each TAP unit was coded with a single coding category, and specific task foci categories were always chosen over non-specific ones whenever this was supported by the data.

5. RESULTS

5.1. Task Foci Distribution

The distribution of all task foci observed in the first run-through of the think-aloud task is presented in Figure 3 for all texts and participants.

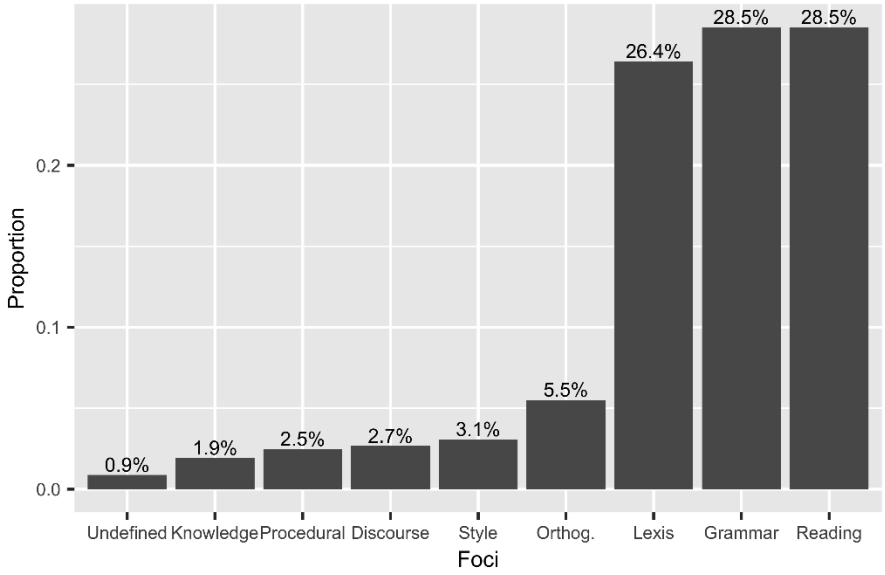

Figure 3
Bar chart with task foci distribution for the first run-through of the think-aloud task, based on all participants and texts; the chart displays rounded percentages

The distribution of coding units across different TAP categories shows that Reading and Grammar were the most frequent task foci, each accounting for approximately 28.5% of the coding units. Lexis was the third most prominent task focus, comprising 26.4% of the coding units. These three TAP categories had 83.4% of the coding units altogether, which means that most of what goes on in post-editors' minds involve non-specific reading/evaluation, or processes related to lexical or grammatical/syntactical aspects of the task.

The results for Lexis are particularly noteworthy. Despite the fact that the MT output provides post-editors with lexical suggestions (i.e. the MT output itself), the prominence of this category indicates that lexis is not a peripheral issue in post-editing. This seems consistent with a pattern that holds for a number of text genres. Previous research in areas as far apart as technical and poetry translation, for example, call attention to the importance of lexis (see Newmark 1988:152; Jones 2011:129). The results reported here suggest that post-editing too is among the kinds of translating activity where lexis plays a central role.

5.2. Different Task Foci and Cognitive Effort

The effort measures collected in the eye-tracking task were used to cluster the study's sentence pairs (ST and MT) into groups expected to require different levels of cognitive effort. The clustering procedure was performed in the Weka toolkit (Hall et al. 2009) with the *K* means algorithm (MacQueen 1967). Per-sentence means of average fixation duration, fixation count (normalised by source character), and subjective ratings on cognitive effort were used to automatically assign the sentences to one of three possible clusters. Relative to each other, the clusters corresponded to "low", "medium", and "high" levels of cognitive effort, as shown in Table 2.[11] Based on a human evaluation reported by Vieira (2016), it was also found that these clusters corresponded to low, medium and high levels of MT quality (see Vieira 2016:137).

Table 2
Per-cluster averages of cognitive effort based on all 19 subjects taking part in the eye-tracking task

	Low Cog. Effort (180 ST words)	**Medium Cog. Effort** (279 ST words)	**High Cog. Effort** (385 ST words)
avg. fixation duration (in sec.)	0.243	0.262	0.293
subjective cog. effort (1-9)	2.19	3.38	4.94
fixation count (per ST character)	0.47	0.73	1.25

These clusters made it possible to check if the task foci distributions observed in the think-aloud task vary as a function of cognitive effort. Table 3 shows task foci distributions and the total number of coding units in each cluster. It is noticeable that the total number of coding units increases together with the level of cognitive effort each cluster is expected to require, which supports the distinction between the clusters also in the think-aloud task. This also counters potential concerns regarding the fact that participants edited the text at a sentence level in the eye-tracking task and at a text level in the think-aloud task. Differences of this kind would be expected to induce discrepancies between the two tasks rather than drive the linear association observed between the cognitive effort clusters and the number of TAP units, so this is not regarded here as a serious issue (see also Vieira 2016:154-155). In addition, despite previous criticisms, this suggests that the think-aloud interference does not invalidate the use of TAPs as a method to estimate cognitive effort, as in the strategy adopted by Krings (2001). Indeed, a detailed analysis of the TAPs obtained in the present study revealed interesting connections between verbalisations and other types of data, but these results are not reported here, as this is beyond the scope of this paper (see Vieira 2016).

Table 3
Incidence of different task foci across sentence clusters expected to pose low, medium, and high levels of cognitive effort; the values between brackets are TAP unit counts normalised by the number of source words in each cluster

	Low units (/ST word)	**Medium** units (/ST word)	**High** units (/ST word)
Lexis	89 (0.5)	345 (1.24)	883 (2.29)
Grammar	82 (0.45)	317 (1.14)	934 (2.43)
Style	29 (0.16)	10 (0.04)	111 (0.29)
Orthography	22 (0.12)	40 (0.14)	180 (0.47)
Discourse	14 (0.08)	30 (0.11)	73 (0.19)
Knowledge	27 (0.15)	14 (0.05)	24 (0.06)
Reading	252 (1.4)	375 (1.34)	733 (1.90)
Procedural	8 (0.04)	12 (0.04)	67 (0.17)
Undefined	2 (0.01)	10 (0.04)	25 (0.06)
Total	*525 (2.91)*	*1154 (4.14)*	*3030 (7.87)*

As regards differences between task foci, it can be noted that Lexis and Grammar increase more sharply from the low-effort to the high-effort cluster. This suggests that these categories are both frequent and cognitively demanding. Reading has a less pronounced pattern: the low- and medium-effort clusters had nearly the same incidence of the Reading

TAP category (relative to cluster size). This result is not surprising, as post-editors are required to read and/or evaluate the text under all circumstances and not only in high-effort conditions.

Regarding the other task foci, it is worth noting that the low-effort cluster has the majority of Knowledge TAP units. Here it should be mentioned that most of these units corresponded to a single sentence, presented in example (10):

(10) ST: *On peut télécharger ce document (en anglais pour l'instant, une traduction sera offerte sous peu) à cette adresse: http://ca.movember.com/fr/mens-health/prostate-cancer-screening* 'You can download this document (in English for the time being, a [French] translation will be available shortly) at this address: http://ca.movember.com/fr/mens-health/prostate-cancer-screening'

MT: *You can download this document (in English for the moment, a translation will be provided shortly) to this address: http://ca.movember.com/fr/mens-health/prostate-cancer-screening*

The passage in example (10) refers to a website that was only available in English. Since the texts were post-edited into English, most participants decided not to provide the information about a French translation and deleted the passage between brackets in the MT output. This is an example of a situation where knowledge of the real-world use of the text can have a direct impact on participants' behaviour, which underlines the importance of taking contextual aspects of this kind into account in empirical investigations of the post-editing process. Other occasions where participants focused on similar issues involved, for example, the use of acronyms, which requires knowledge of how specific terms are worded in different contexts.

Units in the Orthography TAP category were associated mostly with low-quality MT sentences that had malformed or misspelt words. Hyphenation issues were also prominent in this category, but less so than spelling.

When cluster size is controlled for, Style, Knowledge, and Reading have a non-linear pattern across the clusters, which suggests that these task foci do not have a straightforward association with cognitive effort. Aspects other than effort might be more strongly related to these TAP categories. Knowledge, for example, seems highly dependent on the content of the text, while Style might be more directly related to post-editors' own preferences and individual characteristics.

To investigate further the patterns observed in Table 3, mixed-effects regression models (Baayen, Davidson, and Bates 2008) were fitted to the data[12] to measure how

cognitive effort related to different task foci. This was done based on binomial comparisons, i.e. by checking if cognitive effort was more strongly related to the specific task foci or to Reading, which was regarded as a neutral TAP category. Reading seemed a good comparison parameter because, as mentioned earlier, reading/evaluating the text is a requirement of the task rather than a behaviour expected to be associated with a specific level of effort. Mixed-effects modelling allows participants and items (in the present study, post-editors and sentences, respectively) to be treated as random factors.[13] This method controls for effects associated just with the participants or the textual materials sampled for the study, which enhances the generalizability of the findings.

The TAP coding units were the data points in the analysis. The cognitive effort clusters (a three-level categorical variable) and participant variables (namely, level of professional experience, score in the French test, and attitude to MT) were tested as potential predictors. All numeric variables were scaled to mean = 0 and variance = 1. Insignificant variables were removed from the models as per Balling and Baayen (2008) (see also Hvelplund 2011:126).

Table 4
Results in mixed-effects binomial models comparing specific task foci with Reading/Evaluation
* = p < 0.05; ** = p < 0.01; *** = p < 0.001

	Lexis		**Grammar**		**Knowledge**		**Style**		**Discourse**		**Orthography**	
Observations	2677		2693		1428		1510		1477		1602	
	β	z	β	z	β	z	β	z	β	z	β	z
c. effort 2 - 1	1.31	3.25**	0.92	4.41***	-	-	-1.54	-2.71*	-	-	-	-
c. effort 3 - 1	1.78	4.62***	1.40	7.16***	-	-	0.1	0.23†	-	-	-	-
c. effort 3 - 2	0.47	1.45†	0.48	3.1**	-	-	1.64	3.24**	-	-	-	-
experience	0.31	1.99*	0.20	2.27*	0.67	2.49*	-		0.66	3.13**	0.27	2.44*
attitude	-	-	-		-		-0.45	-3.16**	-	-	-	-

Values for each predictor are regression coefficients (β) together with the Wald statistic (z) (higher z-score values stand for lower standard errors) and significance level. Cognitive effort is presented in terms of between-level comparisons. Subjects and items were treated as random effects in all models.
† = non-significant two-way comparisons that were kept in the model, as they are part of a single categorical variable: cognitive effort.

Table 4 presents the variables found to be significant ($p < 0.05$) in each model.[14] Positive coefficients indicate a higher occurrence probability for specific task foci relative to Reading. The Procedural and Undefined TAP categories were not analysed. A higher level of professional experience was related to a higher occurrence probability for TAP units in the

Lexis (β = 0.31), Grammar (β = 0.20), Knowledge (β = 0.67), Discourse (β = 0.66), and Orthography (β = 0.27) categories. By contrast, a negative coefficient is observed for attitude towards MT (β = -0.45) in the Style model. This indicates that most TAP categories polarised participants with respect to level of professional experience, but Style polarised them with respect to attitude towards MT. This has at least two implications. First, it may be that one of the reasons underlying a negative attitude towards the use of MT relates to a low level of tolerance for MT passages with poor style. Second, professional participants may be more capable of focusing on specific linguistic issues during the task, such as lexis and grammar, which is interesting, since previous research (e.g. de Almeida 2013) has found little effect of professional experience in post-editing (though see Moorkens and O'Brien 2015).

The variable representing cognitive effort (i.e. the three clusters, referred to as 'c. effort' in Table 4) was only kept in the models if its overall effect was found to be significant. All pairwise comparisons of this variable were tested,[15] i.e. by checking if the occurrence probability of the TAP units changed from cognitive effort level 1 to 2 ('c. effort 1 - 2'), from level 2 to 3 ('c. effort 3 - 2'), and from level 1 to 3 ('c. effort 3 - 1'). As can be seen, higher levels of cognitive effort are significantly related to a higher occurrence probability for TAP units in the Grammar and Lexis categories. Significant differences were observed between all cognitive effort levels for Grammar (β = 0.48, β = 0.92, β = 1.40), and between levels 1 and 3, and 1 and 2 for Lexis (β = 1.31, β = 1.78). These results match the patterns presented in Table 3, where Grammar and Lexis have a clear association with cognitive effort.

Table 4 also shows a non-linear relationship between cognitive effort and the Style TAP category. The occurrence probability for the Style category significantly decreased from cognitive effort level 1 to level 2 (β = -1.54), and significantly increased from level 2 to level 3 (β = 1.64). Seen widely, Style appears to be associated with a different pattern of behaviour compared to the other TAP categories. In any case, the Style category had more TAP units in the top-effort cluster, which suggests that this category too has some type of association with cognitive effort.

An additional binomial model was used to examine the extent to which Lexis and Grammar differed between each other in how they related to cognitive effort. Here Grammar was associated with a higher level of cognitive effort in comparison with Lexis, but this difference was not found to be significant.[16] This is an interesting finding, as in previous research lexis and syntax were deemed to be at the opposite ends of the post-editing effort spectrum (Temnikova 2010). To the knowledge of the present author, the significance of the gap between lexis and grammar had not been tested to date in post-editing considering

aspects of the task involving the ST and intended use of the translation. Results presented here in this respect suggest that the demands of lexis in post-editing should not be overlooked.

5.3. Unit Sequences

It was observed that TAP units did not necessarily have a 1:1 relationship with certain issues participants came across, so it seemed desirable to experiment with a broader segmentation of the TAP data. For this purpose, the TAPs were grouped into sequences consisting of adjacent TAP units that corresponded to the same text string put in working memory for processing. Jones (2011) followed a similar approach in the analysis of TAPs in translation, but he coded the sequences themselves as opposed to individual TAP units.

The TAP sequences in the present study often involved more than one task foci. This occurred, for instance, when an editing operation related to grammar was immediately followed by further edits concerning lexis, or when an edit and a corresponding comment were deemed to represent different task foci. In sequence (S) 912, presented in example (11), after reading a phrase in the MT output ('the test of APS'), P26 turns her attention to an incorrect acronym in the machine translation (see TAP units 4251 and 4252). After editing the acronym, P26 implements a structural change, opting to use 'PSA' as a modifier. In the present study, all these units belonged to the same TAP sequence, as they all pertained to the same text string that had been put in working memory.

(11)	S912	U4249	I'll have a look at the next one	(Procedural)
		U4250	urm, OK, *the test of APS*	(Reading)
		U4251	so I guess this is referring still to the same test	(Knowledge)
		U4252	so it should be ***PSA*** [~~*APS*~~]	(Knowledge)
		U4253	urm and I'm going to call it the *PSA* ***test*** [~~*test of*~~]	(Grammar)

TAP sequences had 4.5 coding units on average. Table 5 shows the average sequence length (in coding units, with standard deviations) and the total number of sequences per cognitive effort cluster. The number of sequences has a clear linear relationship with the cognitive effort levels: it increases from the low-effort to the high-effort cluster. Average sequence lengths present a similar pattern, but here the values for the clusters are closer together. This suggests that the way in which participants mentally segment the text in post-editing is not strongly related to cognitive effort. To examine if other factors could be linked

to the process of mentally segmenting the text, the cognitive effort clusters and different task foci were tested as potential factors affecting the length of TAP sequences.

Table 5
TAP sequence count and average length (in coding units) for each cognitive effort cluster

	Low Cog. Effort	**Medium Cog. Effort**	**High Cog. Effort**
Avg. no. of TAP units per sequence	3.89 (2.76 SD)	4.42 (2.98 SD)	4.68 (3.32 SD)
Total sequence count (/ST word)	133 (0.74)	263 (0.94)	651 (1.69)

Table 6 shows the percentage of different task foci in the bottom and top quartiles (25%) of TAP sequences (containing short and long sequences, respectively). As can be seen, Lexis and Orthography are the task foci with the largest difference in TAP sequence length between the bottom and top quartiles. The pattern observed for Orthography is unsurprising; it would be unusual for issues in this category (e.g. spelling) to require long TAP sequences. The result observed for the Lexis category is more noteworthy, as lexis usually concerns short textual units (i.e. words and expressions), so it is interesting that the Lexis TAP category seems to be related to longer sequences of TAP units. Equally interesting is the fact that no particular effect is observed for Grammar, which has a very similar incidence of TAP units in the bottom and top quartiles of TAP sequences (31% and 27% respectively).

Table 6
Different task foci in short and long sequences

	Bottom Quartile (short sequences)	**Top Quartile** (long sequences)
Lexis	18%	30%
Grammar	31%	27%
Style	2%	3%
Discourse	6%	2%
Orthography	10%	3%
Knowledge	2%	2%
Reading	26%	29%

To take cognitive effort into account and control for variations across participants and items, a *poisson* mixed-effects model was fitted to the data. The number of TAP units in each

sequence was the model's response variable. Cognitive effort and the ratio of Grammar, Lexis and Orthography units in the sequences (ranging between 0 and 1) were tested as predictors. Potential impacts of the participant variables presented in Table 1 were also examined. The results confirmed the patterns observed in Table 6: Lexis has a significant positive association with TAP sequence length ($\beta = 0.23$, $z = 4.39$, $p < 0.001$) while Orthography has a negative effect ($\beta = -0.41$, $z = 4.20$, $p < 0.001$). Grammar presented no significant relationship with sequence length, and nor did cognitive effort. The results concerning Grammar and Lexis are illustrated in examples (12) and (13).

(12)	S286	U1217	*and twenty per cent of the electorate* <u>*between*</u>	(Grammar)
		U1218	rather than ~~*in*~~	(Grammar)
		U1219	***between*** *eighteen* [~~*to*~~] ***and*** *twenty-nine* [~~*years*~~]	(Grammar)
	S488	U1012	*it is in this spirit that a majority of governments*	(Reading)
		U1013	***American*** *governments*	(Grammar)
		U1014	rather than *governments* ~~*American*~~	(Grammar)
(13)	S346	U1583	[*the important*] *thing is to have a debate with*	(Reading)
		U1584	I think *have a discussion*	(Lexis)
		U1585	rather than *debate*	(Lexis)
		U1586	*have a...*	(Lexis)
		U1587	or *talk* maybe [~~*discussion*~~]	(Lexis)
		U1588	*have a talk*	(Lexis)
		U1589	I'll say ***talk***	(Lexis)
		U1590	sounds better	(Lexis)
		U1591	*have a talk with the doctor to decer* determine if they should pass him or not*	(Reading)

Example (12) has two three-unit sequences showing edits related to prepositions and to word order, respectively. Example (13) shows a nine-unit sequence concerning the replacement of 'debate' with 'talk', i.e. a lexical substitution. These two examples illustrate the patterns observed in Table 6 and in the statistical model: decisions of a lexical nature were related to longer sequences of mental processing, which was not observed for edits involving grammar/syntax. This may be because certain MT grammar errors are obvious (e.g. the wrong position of the adjective in S488) and therefore do not require long deliberations to be corrected. Lexis may be more prone to longer sequences of thought processes because different lexical possibilities seem to be predominantly analysed on the paradigmatic (i.e.

vertical) axis, where mutually exclusive alternatives overlap. When post-editors deal with grammar issues, by contrast, different sequences arise as new text strings are put into working memory, so these issues would be expected to be predominantly analysed on the syntagmatic (i.e. horizontal) axis.

Interestingly, it was noted that trying to think of synonyms for a word without directly considering the immediate context can be an inefficient way of solving lexical problems. The MT output for one of the texts contained the string 'surveillance of the disease'. This was not a very adequate phrase, as it referred to a doctor-patient relationship rather than to wider public health procedures. P25 realised this issue, but failed to find a solution for it the first time round, stating, 'I'm not very keen on *surveillance*, but I can't think of a better word'. In this instance, P25 seemed to focus specifically on 'surveillance', relying exclusively on the paradigmatic axis in trying to think of potential synonyms. P24 also realised this issue; however, P24 found a solution for it almost immediately, stating, 'I think you probably ***monitor*** [~~*surveillance*~~] a disease'. It is interesting no note that P24 regarded the issue from paradigmatic as well as syntagmatic perspectives. That is, she thought of a synonym that would adequately collocate with 'disease'. Only 7 out of the study's 28 participants managed to solve this problem. This shows how easy it is to overlook collocation issues in post-editing, which can be a problem if a product of high quality is required.

6. CONCLUDING REMARKS

TAPs collected in the process of post-editing two texts were analysed here to shed light on the relationship between cognitive effort and the different aspects of the tasks addressed and/or thought of by post-editors. Cognitive effort was approximated with a combination of gaze data and subjective ratings gathered independently from the TAPs, in a task carried out by a different, but comparable, sample of participants. While the study sample is relatively small, the results that emerged from this design lead to three main findings. First, lexis, grammar/syntax, and style were the only task foci (i.e. issues addressed and/or thought of) that had a significant relationship with cognitive effort in post-editing. This was found to be the case especially for grammar and lexis. Second, grammar is more cognitively demanding than lexis, but this difference was not found to be significant. Third, post-editors' decision-making involved longer sequences of thought processes when the issue in question concerned lexis.

While both Grammar and Lexis were found to be linked to cognitive effort, it was surprising that the difference between these categories was not significant. In a previous investigation based on MT errors, 'incorrect word' is listed at level 3 out of 10 on a scale that estimates cognitive effort in post-editing, where 10 is the most demanding level (Temnikova 2010). While it was also observed here that grammar is more strongly related to cognitive effort than lexis (see also Koponen et al. 2012), based on a comprehensive consideration of the mental processes that take place in the task, including processes related to the ST and real-world use of the translation, it was found here that the difference in the cognitive demand of lexis and grammar is not significant. In this respect, it should be noted that participants in the present study were instructed to aim for a post-edited product of high quality. This may have triggered processes where semantically close lexical items are chosen over one another, which is expected in tasks where a product of quality similar or equal to traditional translation is required (see TAUS/CNGL 2010). However, even in a post-editing task where stylistic changes were not recommended, previous research found that MT errors relating to terminology and false cognates have a strong correlation with cognitive effort (Lacruz, Denkowski, and Lavie 2014). This seems to rule out a high product quality expectation as the sole reason behind the results observed here, indicating that the challenges of lexis should not be underestimated in post-editing.

Regarding the link between lexis and the length of thought sequences, it is hypothesised here that this result is due to the intrinsic nature of lexical decisions. This is because mutually exclusive lexical alternatives would be expected to be analysed predominantly on a paradigmatic axis, which is not expected in the case of grammatical/syntactical issues. Another possibility in this respect is that different lexical alternatives may simply exist in larger quantities in the language, resulting in longer thought sequences in the process of making a decision. Irrespective of the reason underlying this phenomenon, it was observed that an over-reliance on the paradigmatic axis can be an inefficient way of dealing with lexical issues in post-editing. This was illustrated with the example that involved 'monitoring' and 'surveillance' as potential collocates for 'disease'. The few participants who successfully replaced 'surveillance', an inadequate term in the context, concentrated on the horizontal relationship between the words in the sentence and not on isolated synonyms on the paradigmatic axis.

Interestingly, participants' deliberation over certain lexical issues involved relatively common words that would not be expected to pose significant problems – e.g. the decision between 'discussion', 'talk', and 'debate', shown in example (13). It is not clear if

participants would have consulted external resources in real-world settings in these cases, but this seems unlikely given the ordinary nature of these words. In any case, examples of this kind suggest that monolingual lexical aids (e.g. thesauri) might have a positive effect on the post-editing process. In light of these findings, placing more emphasis on these resources and exploring their use in post-editor training seems like an interesting direction for future research.

Finally, similarly to a pattern previously observed for translation (Carl, Kay and Jensen 2010), three task phases were identified here, corresponding to the stages of gisting, drafting, and revision. Because of a lack of differences in keyboarding across these stages, previous research suggests that 'these phases are interleaved' in post-editing (Green, Heer, and Manning 2013:447). However, based on the TAPs, results observed here suggest that a distinction between these phases can be made in post-editing as well. It should be noted that in Green, Heer, and Manning (2013) post-editing was carried out sentence by sentence, a setting similar to the one adopted in the eye-tracking task described above. In contexts of this kind, quick editing operations implemented in the short textual span of a sentence may indeed dispense with separate phases for gisting and/or revising the text in post-editing. However, it was shown here that if participants are given a longer text, and are allowed to plan their editing strategy, some of them choose to 'gist' the text in advance and/or to have a final run-through to check their work. In view of this, this article postulates that these phases are not dependent on the type of task itself (i.e. translation or post-editing), but rather on whether or not the task is conducive to a text-level translating/editing approach.

Notes

[1] Provided participants are motivated in investing effort in the task, mental effort can be assumed to reflect mental load (i.e. the difficulty/demands of an exercise or experiment). Paas (1992) refers to these two concepts (mental effort and mental load) and performance in the task (e.g. as measured by the errors made) as 'cognitive load'. As the present study is concerned primarily with amount of allocated effort, the term 'mental effort' is maintained here as per the construct directly referred to in Paas's scale. 'Mental' and 'cognitive' effort are regarded here as interchangeable.

[2] For more information on the study's methodology, see Vieira (2016). Details of the eye-tracking task can also be found in Vieira (2014), where textual features and participants' working memory capacity are contrasted with post-editing effort.

[3] Different systems are ranked as part of this workshop, which takes place every year. The dataset from the 2013 edition is freely available at http://www.statmt.org/wmt13/results.html (accessed 07 November 2014).

[4] Available at: http://www.lapresse.ca/vivre/sante/201211/30/01-4599309-depistage-du-cancer-de-la-prostate-passer-le-test-ou-non.php (accessed 05 November 2014)

[5] Available at: http://www.lapresse.ca/la-tribune/opinions/201207/30/01-4560667-une-strategie-republicaine-pour-contrer-la-reelection-dobama.php (accessed 05 November 2014)

[6] The additional systems were SDL Freetranslation.com (http://www.freetranslation.com), PROMT (http://www.online-translator.com/?External=aspForms&prmtlang=en), TransPerfect

(http://web.transperfect.com/free-translations/) Microsoft Translator, via MS Word, and SDL Automated Translation, via Trados Studio 2011 (all harvested in October 2013). Of these, SDL Freetranslation.com, TransPerfect and Microsoft Translator entered the study sample after the selection procedure.
[7] Further information on the selected materials see Vieira (2014), where a different analysis was carried out using part of the dataset obtained in the eye-tracking task.
[8] Using the Tobii VT-I fixation filter, set to default preferences except for an option that discarded all individual fixations below 100 milliseconds (Hvelplund 2011). Data points where the mean duration of fixations was below 200 milliseconds (Hvelplund 2011) were not considered in the analysis. A further seven data points were excluded because total fixation duration was deemed too low (less than 29% of total editing time, which was more than 2.5 standard deviations from the sample mean).
[9] The ethics committee at Newcastle University, where the tasks were conducted, approved the recruitment of human participants for the study.
[10] French Vocabulary: W = 65.5, p = 0.3; experience: W = 75.5, p = 0.6; attitude: W = 102.5, p = 0.4; age: W = 67.5, p = 0.39.
[11] It was checked if excluding participants with low scores in the French test (below 70.4, the average for the eye-tracking sample) would significantly affect the cluster means, but this was not found to be the case; the clusters still corresponded to low, medium, and high levels of effort in relation to each other.
[12] By the Laplace approximation, using the `glmer` function of the `lme4` R package (Bates, Maechler and Bolker 2012).
[13] Random variables are those that do not have a specific range or number of levels, having been sampled from a larger population of unknown size.
[14] Significance was calculated with the `summary` function.
[15] A larger number of significance tests inflates the chance of false positives. Tukey's test was used here to compensate for this inflated risk.
[16] Non-significant results in this respect are not presented here for economy of space.

Acknowledgements

This research was supported by Newcastle University. The author would like to thank Dr Francis Jones, Dr Michael Jin, and Dr Ya-Yun Chen for their comments on an earlier draft of this article.

References

Alves, F., Koglin, A., Mesa-Lao, B., Martínez, M. G., Fonseca, N. B. L., Sá, A. M., Gonçalves, J. L., Szpak, K. S., Sekino, K., & Aquino, M. 2016. Analysing the Impact of Interactive Machine Translation on Post-editing Effort. In: Carl, M., Bangalore, S. & Schaeffer, M. (eds) *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB.* Heidelberg: Springer. 77-94.

Aziz, W., Koponen, M. & Specia, L. 2014. Sub-Sentence Level Analysis of Machine Translation Post‑Editing Effort. In: O'Brien, S., Balling, L.W., Carl, M., Simard, M. & Specia, L. (eds) Post-Editing of Machine Translation: Processes and Applications. Newcastle upon Tyne: Cambridge Scholars Publishing. 170–199.

Aziz, W., Castilho, S. & Specia, L. 2012. *PET: A Tool for Post-Editing and Assessing Machine Translation*. Paper presented at the 8th International Conference on Language Resources and Evaluation, 21-27 May, 2012, Istanbul, Turkey.

Baayen, R. H. 2008. *Analysing Linguistic Data: A Practical Introduction to Statistics*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J. & Douglas Bates, M. 2008. Mixed-effects Modeling With Crossed Random Effects for Subjects and Items. *Journal of Memory and Language* Vol. 59. No. 4.

390-412.

Balling, L. W. & Baayen, R. H. 2008. Morphological Effects in Auditory Word Recognition: Evidence from Danish. *Language and Cognitive Processes* Vol. 23. No. 7-8. 1159-1190.

Bates, D., Maechler, M. & Bolker, B. 2012. *lme4: Linear Mixed-effects Models Using S4 Classes. R Package Version 0.999999-0* [computer software]. Retrieved from http://CRAN.R-project. org/package0lme4.

Carl, M. 2012. *Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research*. Paper presented at the 8th International Conference on Language Resources and Evaluation, 21-27 May, 2012, Istanbul, Turkey.

Carl, M., Kay, M. & Jensen, K. T. H. 2010. *Long Distance Revisions in Drafting and Post-Editing*. Paper presented at the 11th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2010, Iaşi, Romania, March 21–27, 2010.

Creswell, J. W. 2009. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Los Angeles: Sage.

de Almeida, G. 2013. *Translating the Post-Editor: An Investigation on Post-Editing Changes and Correlations with Professional Experience across two Romance Languages*. PhD Thesis. Dublin: Dublin City University.

Denkowski, M. & Lavie, A. 2011. *Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems*. Paper presented at the 6th Workshop on Statistical Machine Translation (EMNLP 2011), July 30-31, 2011, Edinburgh, UK.

Ericsson, K. A. & Simon, H. A. 1980. Verbal Reports as Data. *Psychological Review* Vol. 87. No. 3. 215-251.

Green, S., Heer, J. & Manning, C. D. 2013. *The Efficacy of Human Post-Editing for Language Translation*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, 27 April - 2 May, 2013, Paris, France.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009. The WEKA Data Mining Software: an Update. *SIGKDD Explorations* Vol. 11. No. 1. 10-18.

Hvelplund, K. T. 2011. *Allocation of Cognitive Resources in Translation: An Eye-tracking and Key-logging Study*. PhD Thesis. Copenhagen: Copenhagen Business School.

Jakobsen, A. L. 2003. Effects of Think Aloud on Translation Speed, Revision and Segmentation. In: Alves, F. (ed) *Triangulating Translation. Perspectives in Process Oriented Research.* Amsterdam: Benjamins. 69-95.

Jones, F. R. 2011. *Poetry Translating as Expert Action: Processes, Priorities and Networks*. Amsterdam: John Benjamins.

Just, M. & Carpenter, P. 1980. A Theory of Reading: from Eye Fixation to Comprehension. *Psychological Review* Vol. 87. No. 4. 329-354.

Koehn, P., Alabau, V., Carl, M., Casacuberta, F., García-Martínez, M., González-Rubio, J., Keller, F., Ortiz-Martínez, D., Sanchis-Trilles, G. & Germann, U. 2015. CASMACAT: Final Public Report. Accessed 02 May 2017. http://www.casmacat.eu/uploads/Deliverables/final-public-report.pdf.

Koponen, M. 2012. Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations. Paper presented at the Seventh Workshop on Statistical Machine Translation, June 7-8, Montréal, Canada.

Koponen, M., Aziz, W., Ramos, L. & Specia, L. 2012. *Post-editing Time as a Measure of Cognitive Effort*. Paper presented at the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), October 28, 2012, San Diego, USA.

Krings, H. P. 2001. *Repairing Texts : Empirical Investigations of Machine Translation Post-Editing Processes*. Kent, Ohio: Kent State University Press.

Lacruz, I., Denkowski, M. & Lavie, A. 2014. *Cognitive Demand and Cognitive Effort in Post-Editing*. Paper presented at the 11th Conference of the Association for Machine Translation in the Americas – Third Workshop on Post-EditingTechnology and Practice, 22-26 October, 2014, Vancouver BC, Canada.

Landis, J. R. & Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* Vol. 33. 159-174.

MacQueen, J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, California: University of Calfornia Press. 281-297.

Meara, P. & Buxton, B. 1987. An alternative to Multiple Choice Vocabulary Tests. *Language Testing* Vol. 4. No. 2. 142-154.

Moray, N. 1967. Where is Capacity Limited? A Survey and a Model. *Acta Psychologica* Vol. 27. 84-92.

Moorkens, J. & O'Brien, S. 2015. Post-Editing Evaluations: Trade-Offs between Novice and Professional Participants. In: El-Kahlout, I. D., Özkan, M., Sánchez-Martínez, F., Ramírez-Sánchez, G., Hollowood, F., Way, A. (eds) Proceedings of the 18th Annual Conference of the European Association for Machine Translation (11-13 May, 2015, Antalya, Turkey) Stroudsburg: Association for Computational Linguistics. 75–81.

Newmark, P. 1988. *A Textbook of Translation*. Hertfordshire: Prentice Hall International.

O'Brien, S. 2011. Towards Predicting Post-Editing Productivity. *Machine Translation* Vol. 25. No. 3. 197-215.

O'Donnell, R. D. & Eggemeier, F. T. 1986. Workload Assessment Methodology. In: Boff, K. R., Kaufman, L. and Thomas, J. P. (eds) *Handbook of Perception and Human Performance*. New York: John Wiley and Sons. 42-1-42-49.

Paas, F. G. 1992. Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *Journal of Educational Psychology* Vol. 84. No. 4. 429-434.

Paas, F., Tuovinen, J. E., Tabbers, H. & van Gerven, P. W. M. 2003. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. Educational Psychologist Vol. 38 No. 1. 63-71.

Read, J. 2007. Second Language Vocabulary Assessment: Current Practices and New Directions. *International Journal of English Studies* Vol. 7 No. 2. 105-126.

Specia, L. 2011. *Exploiting Objective Annotations for Measuring Translation Post-Editing Effort*. Paper presented at the 15th International Conference of the European Association for

Machine Translation, 30-31 May, 2011, Leuven, Belgium.

Sun, S. 2011. Think-Aloud-Based Translation Process Research: Some Methodological Considerations. *Meta* Vol. 56. No. 4. 928-951.

TAUS/CNGL. 2010. *Machine Translation Postediting Guidelines*. Accessed 7 November 2014. https://evaluation.taus.net/resources/guidelines/post-editing/machine-translation-post-editing-guidelines.

Temnikova, I. 2010. *A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment*. Paper presented at the 7th International Conference on Language Resources and Evaluation, 17-23 May, 2010, Valetta, Malta.

Tyler, S. W., Hertel, P. T., McCallum, M. C. & Hellis, H. C. 1979. Cognitive Effort and Memory. *Journal of Experimental Psychology: Human Learning and Memory* Vol. 5. No. 6. 607-617.

Vieira, L.N. 2014. Indices of Cognitive Effort in Machine Translation Post-Editing. *Machine Translation*. Vol. 28. No. 3-4. 187-216.

Vieira, L.N. 2016. *Cognitive Effort in Post-Editing of Machine Translation: Evidence from Eye Movements, Subjective Ratings, and Think-Aloud Protocols.* PhD Thesis. Newcastle upon Tyne: Newcastle University.